A Test for Endogeneity in Conditional Quantiles

Tae-Hwan Kim, Christophe Muller

amse
aix marseille
school of economics

Working Papers / Documents de travail

A Test for Endogeneity in Conditional Quantiles

Tae-Hwan Kim
Christophe Muller

WP 2013 - Nr 42

A Test for Endogeneity in Conditional Quantiles

Tae-Hwan Kim[a] and Christophe Muller[b,*]

[a]School of Economics, Yonsei University, Seoul 120-749, Korea.
Tel.: +82-2-2123-5461; fax: +82-2-2123-8638.
tae-hwan.kim@yonsei.ac.kr

[b]Aix-Marseille University (Aix-Marseille School of Economics), CNRS and EHESS,
14, avenue Jules Ferry, F-13621 Aix-en-Provence Cedex, France.
christophe.muller@univ-amu.fr

August 2013

Abstract: In this paper, we develop a test to detect the presence of endogeneity in conditional quantiles. Our test is a Hausman-type test based on the distance between two estimators, of which one is consistent only under no endogeneity while the other is consistent regardless of the presence of endogeneity in conditional quantile models. We derive the asymptotic distribution of the test statistic under the null hypothesis of no endogeneity. The finite sample properties of the test are investigated through Monte Carlo simulations, and it is found that the test shows good size and power properties in finite samples. As opposed to the test based on the IVQR estimator of Chernozhukov and Hansen (2006) in the case of more than a couple of variables, our approach does not imply an infeasible computation time. Finally, we apply our approach to test for endogeneity in conditional quantile models for estimating Engel curves using UK consumption and expenditure data. The pattern of endogeneity in the Engel curve is found to vary substantially across quantiles.

Key words: regression quantile, endogeneity, two-stage estimation, Hausman test, Engel curve.

JEL codes: C21.

*The first author is grateful for financial support from the National Research Foundation of Korea – a grant funded by the Korean Government (NRF-2009-327-B00088).

1 Introduction

The issue of endogeneity in the context of quantile regression has long been recognized, and many techniques to deal with it have been proposed.[1] However, not much attention has been paid to the issue of testing for the presence of endogeneity in conditional quantile models. It has been implicitly assumed in much of the previous literature that the presence of endogeneity in either the conditional mean or a particular conditional quantile implies that the entire conditional distribution (i.e., all other conditional quantiles) is contaminated by endogeneity. Such a restriction appears to be unnecessarily strong. It is more general, and perhaps more realistic, to allow for the possibility that endogeneity is only present in some part of the conditional distribution.

For example, consider a typical wage equation where the logarithm of the wage rate of a worker is linearly explained by education level and some other explanatory factors with constant coefficients. The latter factors are often considered to be independent of the error term. In contrast, the independence of the education variable and the error is generally disputed, particularly because some unobservable genetic ability may be simultaneously related to both wage and education. In that case, the model may be subject to the endogeneity problem.

Moreover, in some contexts the poor may have only limited access to secondary school, perhaps because it is costly, while both the poor and the non-poor are covered by mandatory and free universal primary education. Let us simplify the reasoning by assuming that there are only two education levels: primary and secondary. In that case, if an unobservable ability is useful only for skills learned at secondary school, this endogenous genetic ability may affect the estimation problem only for the non-poor. Given that the non-poor mostly correspond to high quantiles in the quantile model for the wage equation, it seems intuitive that this type of endogeneity is likely to be present mostly in high quantiles in this hypothetical context. Knowing which quantiles are affected by endogeneity would help increase the efficiency of the estimation for the other quantiles because there is no need to introduce efficiency loss due to instrumentation.

As another example, consider a population of car drivers that can be divided into unobservable high and low risk drivers. Assume that each driver's risk characteristics are incorporated in the error term of a linear equation describing his or her insurance premium according to some observable characteristics (again with constant coefficients, as this is the usual practice). Suppose also that this equation includes as a regressor a variable measuring the observed driving skills (e.g., records of fines for excessive speed). Clearly, the latter variable may be endogenous, given that high risks should be associated with bad driving habits. In addition, a lower proportion of high-risk drivers than low-risk drivers may have access to insurance contracts. Such selectivity may affect the endogeneity of the quality variable in the insurance premium equation. This would be the case, for example, if the insurance companies implemented a certain selection policy that purged most of the endogeneity of high risks from the data generation process. In that case, one would expect that, in the corresponding quantile regressions, high quantiles may correspond to exogeneity of the driving quality variable, while low quantiles may still be affected by severe endogeneity. As before, some knowledge of the quantiles affected by endogeneity may allow efficiency gains in the estimation. In that case, the insurance companies would be better able to discriminate between good and bad risks.

The typical Hausman test of endogeneity in linear models is a test of the comparison of OLS estimates with 2SLS estimates (Hausman, 1978). However, it is well known that the mean regression

[1] To name just a few, see Amemiya (1982), Powell (1983), Chen and Portnoy (1996), Kemp (1999), Sakata (2007), Arias et. al. (2001), Garcia et. al. (2001), Chen, Linton, and van Keilegem (2003), Hong and Tamer (2003), Kim and Muller (2004, 2012), Chernozhukov and Hansen (2005, 2006, 2008), Ma and Koenker (2006), Horowitz and Lee (2007), and Lee (2007).

can be seen as the average of quantile regressions. Thus, the gap between OLS and 2SLS may be considered as a consequence of a more general endogeneity issue at diverse quantiles rather than as a complete basis for a test of an often complex endogeneity situation. A Hausman-type test is based on the quadratic distance between two distinct estimators of the same quantile parameters. In such a setting, one estimator is consistent only when there is no endogeneity, while the other remains consistent regardless of the presence of endogeneity.

Developments in Hausman-type tests used for endogeneity analysis have attracted interest in the recent literature.[2] We contribute to this interest by exploring the behaviour of one such test across the quantile set. Furthermore, other authors have investigated additional features of quantile regression under endogeneity by looking across the quantile set, such as instrumenting to achieve identification for different quantiles (e.g. Jun, 2008). Chernozhukov and Hansen (2006) propose an exogeneity test for instrumental quantile regression, which considers the whole quantile process as a whole. A potential drawback of their approach is that the computation times of the test statistics and of its critical value are huge for more than a few endogenous regressors.

There are other less obvious reasons to use quantile regressions to investigate exogeneity issues. In particular, recall that the order statistic is a sufficient statistic under iid. This fact suggests that, in that case, all of the information useful for the test can be reached by using quantile regressions that provide a convenient handle on these order statistics. Because the useful information is optimally captured in such cases, at least for nonparametric quantile regressions, one may hope to be able to construct a test that would dominate other endogeneity tests.

Finally, the issues of weak endogeneity in linear simultaneous systems are often originated in non-normality issues. As it happens, quantile regressions are often motivated by situations believed to be far from the normality hypothesis, where they can be preferable to least-square estimators that are efficient exclusively under normality of the errors. In such situations, if one is interested in endogeneity problems, quantile regressions constitute an interesting alternative investigation tool to mean regression, even when the main interest is in central tendency responses.

In this paper, we propose a formal test for the presence of endogeneity at each given conditional quantile level separately. As usual, our test statistic is based on the distance between two estimators. The first estimator included in our test statistics is the standard quantile regression estimator, and the second estimator is the double-stage quantile estimator developed in Kim and Muller (2004).

We present the model and discuss how it is estimated in Section 2. The proposed test statistic is discussed and its asymptotic distribution is derived in Section 3. The finite sample properties of our test are studied through Monte Carlo simulations in Section 4. In Section 5, we provide comparative computation times for the tests respectively based on DSQR and IVQR approaches. In Section 6, we apply our test to the estimation of Engel curves using UK consumption expenditure data. Finally, Section 7 presents concluding remarks.

2 The Model and the Estimation Method

We are interested in the parameter (α_0) in the following structural equation for T observations:

$$\begin{aligned} y_t &= x_{1t}'\beta_0 + Y_t'\gamma_0 + u_t \\ &= Z_t'\alpha_0 + u_t, \end{aligned} \tag{1}$$

where $[y_t, Y_t']$ is a $(G+1)$ row vector of possible endogenous variables, x_{1t}' is a K_1 row vector of exogenous variables, $Z_t = [x_{1t}', Y_t']'$, $\alpha_0 = [\beta_0', \gamma_0']'$ and u_t is an error term. We denote by x_{2t}' the

[2] Hahn and Hausman (2002), Butler (2000), Chmelarova and Hill (2010), Lee and Okui (2012).

row vector of $K_2(= K - K_1)$ exogenous variables absent from (1).

Estimating α_0 at the θ^{th}-conditional distribution can be achieved through the following minimization program:

$$\min_{\alpha} \sum_{t=1}^{T} \rho_\theta(y_t - Z_t'\alpha) \tag{2}$$

and $\rho_\theta(z) = z\psi_\theta(z)$ where $\psi_\theta(z) = \theta - 1_{[z\leq 0]}$ and $1_{[.]}$ is the Kronecker index. The solution of (2), denoted by $\tilde{\alpha}$, will be called the one-stage quantile estimator for α_0. The one-stage estimator $\tilde{\alpha}$ is consistent if the following zero conditional expectation condition holds:

$$E(\psi_\theta(u_t)|Z_t) = 0. \tag{3}$$

This condition is the assumption that zero is the given θ^{th}-quantile of the conditional distribution of u_t. It identifies the coefficients of the model. However, the condition in (3) is generally violated if there is endogeneity in Y_t, and this problem can be appropriately defined as corresponding to $E(\psi_\theta(u_t)|Z_t) \neq 0$. In this case, $\tilde{\alpha}$ is inconsistent, and a two-stage estimation method can be employed to obtain a consistent estimator. In this paper, we develop a procedure to test for endogeneity in Y_t at each quantile θ.

We assume that Y_t can be linearly predicted from the exogenous variables:

$$Y_t' = x_t'\Pi_0 + V_t', \tag{4}$$

where $x_t' = [x_{1t}', x_{2t}']$ is a K-row vector, Π_0 is a $K \times G$ matrix of unknown parameters, and V_t' is a G row vector of unknown error terms. By assumption, the first element of x_{1t} is 1. Using (1) and (4), y_t can also be expressed as follows:

$$y_t = x_t'\pi_0 + v_t, \tag{5}$$

where

$$\begin{aligned} \pi_0 &= H(\Pi_0)\alpha_0 \text{ with } H(\Pi_0) = \left[\begin{pmatrix} I_{K_1} \\ 0 \end{pmatrix}, \Pi_0\right] \\ \text{and } v_t &= u_t + V_t'\gamma_0. \end{aligned} \tag{6}$$

As mentioned before, our test statistic is based on the double-stage quantile regression in Kim and Muller (2004). The use of this estimator has several advantages over other approaches. First, the calculus involved in simultaneously comparing the asymptotic representations of the two estimators is tractable. Moreover, there is no need for numerical solutions or nonparametric estimation of the models, as is the case with most methods of two-stage quantile regression in the literature. This advantage is important because it avoids the need for grid search and the curse of dimensionality, which both limit the analysis to models with only a few variables because of the computational burden and slow convergence issues.

The equations in (4) and (5) are the basis of the first-stage estimation that yields the consistent estimators $\hat{\pi}$, $\hat{\Pi}$, respectively, of π_0, and Π_0. More specifically, $\hat{\pi}$ and $\hat{\Pi}_j$ (the j^{th} column of $\hat{\Pi}$; $j = 1, \ldots, G$) are first stage estimators obtained by

$$\min_{\pi} \sum_{t=1}^{T} \rho_\theta(y_t - x_t'\pi) \tag{7}$$

$$\text{and} \qquad \min_{\Pi_j} \sum_{t=1}^{T} \rho_\theta(Y_{jt} - x_t'\Pi_j), \tag{8}$$

where π and Π_j are $K \times 1$ vectors and Y_{jt} is the $(j,t)^{th}$ element of Y. Estimating π will be useful later on for providing an estimate of the residual $\hat{v}_t$, which is a component of the estimator of the variance-covariance matrix intervening in the test statistics. Note that another quantile index,θ' that is different from θ could also be chosen for estimating at the first stage. However, because we have no compelling reason for this choices, and to alleviate notations, we keep the same θ. Based on these first-stage estimators, the second-stage estimator $\hat{\alpha}$ is obtained as follows:

$$\min_{\alpha} \sum_{t=1}^{T} \rho_\theta(y_t - x_t' H(\hat{\Pi})\alpha).$$

The resulting estimator $\hat{\alpha}$ is denoted as the double-stage quantile estimator for α_0. In order to derive the asymptotic distributions of $\tilde{\alpha}$ and $\hat{\alpha}$, we impose the following regularity conditions. Let $h(\cdot|x)$, $f(\cdot|x)$, and $g_j(\cdot|x)$ be the conditional densities, respectively, for u_t, v_t, and V_{jt}.

Assumption 1 *The sequence* $\{(Y_t', x_t', u_t, v_t, V_t')\}$ *is independent and identically distributed (iid).*

Assumption 1 is imposed to ease the exposition of our results. It arises, for example, when the considered sources of uncertainty in the data come from sampling randomly the observations. Moreover, the assumed conditions can be relaxed to include serial correlation and heteroskedasticity.

Assumption 2 (i) $E(||x_t||^3) < \infty$ *and* $E(||Y_t||^3) < \infty$ *where* $||a|| = (a'a)^{1/2}$.
(ii) $H(\Pi_0)$ *is of full column rank.*
(iii) *There is no hetero-altitudinality:* $h(\cdot|x) = h(\cdot)$, $f(\cdot|x) = f(\cdot)$ *and* $g_j(\cdot|x) = g_j(\cdot)$ *where* $h(\cdot)$, $f(\cdot)$ *and* $g_j(\cdot)$ *are assumed to be continuous. Moreover, all densities are positive when evaluated at zero:* $h(0) > 0$, $f(0) > 0$, *and* $g_j(0) > 0$.
(iv) *All densities are bounded above; i.e., there exist constants* λ_h, λ_f, *and* λ_j *such that* $h(\cdot) < \lambda_h$, $f(\cdot) < \lambda_f$, *and* $g_j(\cdot) < \lambda_j$.
(v) *The matrices* $Q_x = E(x_t x_t')$ *and* $Q_z = E(Z_t Z_t')$ *are finite and positive definite.*
(vi) $E\{\psi_\theta(v_t) \mid x_t\} = 0$ *and* $E\{\psi_\theta(V_{jt}) \mid x_t\} = 0$ $(j = 1, \ldots, G)$.

Assumption 2(i), the moment condition on the exogenous variables, is necessary for the stochastic equicontinuity of our empirical process in the dependent case, which is used for the asymptotic representation. It is also used to bound the asymptotic covariance matrix of the parameter estimators. Assumption 2(ii) is analog to the usual identification condition for simultaneous equations models. Assumption 2(iii) allows us to simplify the asymptotic covariance matrix of the double-stage estimator. Otherwise, the covariance has a complicated form as shown in Kim and Muller (2004). Assumption 2(iv) simplifies the demonstration of convergence of the remainder terms to zero for the calculation of the asymptotic representation. Assumption 2(v) is the counterpart of the usual condition for OLS under which the sample second moment matrix of the regressor vectors converges towards a finite positive definite matrix. It ensures that $E(x_t Y_t) \neq 0$ and $E(Z_t Y_t) \neq 0$. Lastly, Assumption 2(vi) is the assumption that zero is the θ^{th}-quantile of the conditional distribution of v_t and of each V_{jt}.[3] It identifies the coefficients of the model.

[3]Note that in the iid case, the term $f(F^{-1}(\theta))^{-1}$ typically appears in the variance formula of a quantile estimator (Koenker and Bassett, 1978). However, due to Assumption 3(iv), $F^{-1}(\theta)$ is now zero so that in this case, we instead have $f(0)^{-1}$.

The asymptotic representation of the one-step quantile estimator $\tilde{\alpha}$ is well known:

$$T^{1/2}(\tilde{\alpha} - \alpha_0) = Q_z^{-1} T^{-1/2} \sum_{t=1}^{T} Z_t \epsilon_{1t} + o_p(1), \tag{9}$$

where $\epsilon_{1t} = h(0)^{-1}\psi_\theta(u_t)$. From (9), it is easily seen that $\tilde{\alpha}$ is consistent if $T^{-1}\sum_{t=1}^{T} Z_t\epsilon_{1t}$ vanishes in probability. Given that the probability limit $E(Z_t\epsilon_{1t})$ is zero in the absence of endogeneity, we have in that case

$$T^{1/2}(\tilde{\alpha} - \alpha_0) \xrightarrow{d} N(0, \sigma_{11}Q_z^{-1}),$$

where $\sigma_{11} = E(\epsilon_{1t}^2) = h(0)^{-2}\theta(1-\theta)$ and $Q_z = E(Z_tZ_t')$. The covariance estimator $\sigma_{11}Q_z^{-1}$ can be consistently estimated by $\hat{\sigma}_{11}\hat{Q}_z^{-1}$, where $\hat{Q}_z = T^{-1}\sum_{t=1}^{T} Z_tZ_t'$ and $\hat{\sigma}_{11} = T^{-1/2}\sum_{t=1}^{T}\hat{\epsilon}_{1t}^2 = \hat{h}(0)^{-2}\theta(1-\theta)$ with $\hat{\epsilon}_{1t} = \hat{h}(0)^{-1}\psi_\theta(\hat{u}_t), \hat{u}_t = y_t - Z_t\hat{\alpha}$. Here, $\hat{h}(0)$ can be any consistent kernel-type non-parametric estimator of density h at zero.

A similar result can be obtained for the second-stage estimator $\hat{\alpha}$ (see Kim and Muller, 2004, for more details):

$$T^{1/2}(\hat{\alpha} - \alpha_0) = Q_{zz}^{-1} H(\Pi_0)' T^{-1/2} \sum_{t=1}^{T} x_t\epsilon_{2t} + o_p(1), \tag{10}$$

where $Q_{zz} = H(\Pi_0)'Q_xH(\Pi_0)$, $Q_x = E(x_tx_t')$, and $\epsilon_{2t} = f(0)^{-1}\psi_\theta(v_t) - \sum_{i=1}^{G}\gamma_{0i}g_i(0)^{-1}\psi_\theta(V_{it})$. Therefore, we have the following result:

$$T^{1/2}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, \sigma_{22}Q_{zz}^{-1}),$$

where $\sigma_{22} = E(\epsilon_{2t}^2)$. As before, σ_{22} and Q_{zz} can be consistently estimated: $\hat{Q}_{zz} = H(\hat{\Pi})'\hat{Q}_xH(\hat{\Pi})$ with $\hat{Q}_x = T^{-1}\sum_{t=1}^{T} x_tx_t'$ and $\hat{\sigma}_{22} = T^{-1/2}\sum_{t=1}^{T}\hat{\epsilon}_{2t}^2$ with $\hat{\epsilon}_{2t} = \hat{f}(0)^{-1}\psi_\theta(\hat{v}_t) - \sum_{i=1}^{G}\hat{\gamma}_{0i}\hat{g}_i(0)^{-1}\psi_\theta(\hat{V}_{it})$ where $\hat{f}(0)$ and $\hat{g}_i(0)$ are kernel-type estimators of $f(0)$ and $g_i(0)$, respectively, and $\hat{v}_t$ and $\hat{V}_{it}$ are the residuals from the first-stage regressions in (7) and (8).

3 The Endogeneity Test

The null hypothesis we wish to test is

$$H_0 : \text{There is no endogeneity in the } \theta^{th} \text{ quantile,}$$

which is equivalent to

$$H_0 : E(\psi_\theta(u_t)|Z_t) = 0 \text{ for a given } \theta. \tag{11}$$

Examining this relationship more closely will allow us to discuss how endogeneity at different quantiles can be understood in a similar manner to what is done for the exogeneity notion typically used in LS estimation. Equation (11) for a given θ implies that $E\left\{Y_t(\theta - I_{[u_t<0]})\right\} = 0$, if we believe that the only variable possibly carrying an endogeneity problem is Y_t. For non-centered Y_t, this is equivalent to $\theta = \frac{E(Y_ts_t)}{EY_t}$, where s_t denotes the sign index variable $I_{[u_t<0]}$.

Let us now normalize Y_t as $n_t = Y_t/(EY_t)$. Then we have $\theta = E(n_ts_t|Z_t)$, which implies that $cov(n_t, s_t|Z_t) = \theta - E(s_t|Z_t)$. We note that $E(s_t|Z_t)$ is the population conditional proportion of errors below zero, which is equal to θ as implied by (11). Therefore, under exogeneity, we have $cov(n_t, s_t|Z_t) = 0$, which implies that $cov(Y_t, s_t|Z_t) = 0$. Hence, we have shown that the quantile exogeneity condition can be interpreted as a linear orthogonality condition of the possibly endogenous variables with the sign index s_t. This interpretation is helpful because intuitive reasonings

with regard to the usual zero covariance condition can be used to select instruments in quantile regression in a similar manner to the typical approach used for 2SLS.

The principle driving the test is that the slope coefficients estimated both by $\tilde{\alpha}$ and $\hat{\alpha}$ are consistent for the true value α_0 and asymptotically normal under the null hypothesis of no endogeneity, while only the slope coefficients estimated by $\hat{\alpha}$ are consistent to those of α_0 under the alternative hypothesis. That is: we exclude the intercept in the comparison of the estimators. Thus, a quadratic distance between $\tilde{\alpha}$ and $\hat{\alpha}$, excluding the intercept coefficients, can be used to test consistently the null hypothesis of no endogeneity. If we wanted to place ourselves in the original Hausman test framework (Hausman, 1978), stricto sensu, $\hat{\alpha}$ should be efficient under H_0. However, we cannot use the difference of asymptotic variance-covariance matrices as in the usual Hausman test because quantile regression is not asymptotically efficient even under exogeneity. As a consequence, we need to calculate the covariance of the two estimators, which will be given by the covariance of their asymptotic representations. Thus, we allow for inefficient estimators by dealing with the joint distribution without invoking orthogonality conditions between estimators.

Under the null hypothesis of exogeneity at the θ^{th} quantile, both the quantile regression and the double quantile regression converge to the same values *for the slopes*. On the other hand, under the alternative hypothesis of endogeneity at the at the θ^{th} quantile, the difference of the two slope estimators diverge. These features ensure that our test is consistent.

A simple specification is that of a quadratic distance between $\tilde{\alpha}$ and $\hat{\alpha}$, weighed by the variance matrix of $\tilde{\alpha} - \hat{\alpha}$. It will be shown below that the variance-covariance matrix of $\tilde{\alpha} - \hat{\alpha}$ is given by $RC^{-1}R'$, where $R = \left[\begin{array}{cc} I_{K_1+G} : & -I_{K_1+G} \end{array} \right]$ and C is defined in Lemma 1 below. Hence, a preliminary and ancillary statistic is defined as $T(\tilde{\alpha}-\hat{\alpha})[RC^{-1}R']^{-1}(\tilde{\alpha}-\hat{\alpha})$. We modify this statistic later on to account for possible inconsistent intercept estimators. For the time being, we analyse this statistic as a useful intermediate in the derivation of the final test statistics. The following lemma shows the null distribution of our preliminary statistic. The proof of the lemma is provided in the Appendix.

Lemma 1. Suppose that Assumptions 1 and 2 hold. Then, under the null hypothesis of no endogeneity at quantile θ, we have:

$$T(\tilde{\alpha} - \hat{\alpha})[RC^{-1}R']^{-1}(\tilde{\alpha} - \hat{\alpha}) \xrightarrow{d} \chi^2(K_1 + G),$$

where

$$\begin{aligned} C &= \left[\begin{array}{cc} \sigma_{11} Q_z^{-1} & \sigma_{12} Q_z^{-1} Q_{zx} H(\Pi_0) Q_{zz}^{-1} \\ \sigma_{12} Q_{zz}^{-1} H(\Pi_0)' Q_{zx}' Q_z^{-1} & \sigma_{22} Q_{zz}^{-1} \end{array} \right] \\ \text{and } Q_{zx} &= E(Z_t x_t') \text{ and } \sigma_{12} = E(\epsilon_{1t}\epsilon_{2t}). \end{aligned}$$

For practical implementation, C can be replaced with a consistent estimator $\hat{C}_T$ without affecting the limiting distribution. We can use the plug-in principle to propose the following consistent estimator for C:

$$\hat{C}_T = \left[\begin{array}{cc} \hat{\sigma}_{11} \hat{Q}_z^{-1} & \hat{\sigma}_{12} \hat{Q}_z^{-1} \hat{Q}_{zx} H(\hat{\Pi}) \hat{Q}_{zz}^{-1} \\ \hat{\sigma}_{12} \hat{Q}_{zz}^{-1} H(\hat{\Pi})' \hat{Q}_{zx}' \hat{Q}_z^{-1} & \hat{\sigma}_{22} \hat{Q}_{zz}^{-1} \end{array} \right],$$

where

$$\hat{Q}_{zx} = T^{-1}\sum_{t=1}^{T} Z_t x_t',$$

$$\hat{\sigma}_{12} = T^{-1}\sum_{t=1}^{T} \hat{\epsilon}_{1t}\hat{\epsilon}_{2t}.$$

The consistency of $\hat{C}_T$ is stated in Lemma 2, of which the proof is in the Appendix.

Lemma 2. Suppose that the kernel-type density estimators $\hat{h}(0)$, $\hat{f}(0)$, and $\hat{g}_i(0)$ are respectively consistent for $h(0)$, $f(0)$ and $g_i(0)$, $i = 1, ..., G$. Then, under Assumptions 1 and 2, we have

$$\hat{C}_T \xrightarrow{p} C.$$

We now deal with the fact that the intercept estimator in $\hat{\alpha}$ may not be consistent for all values of θ. Because the semi-parametric restrictions $E\{\psi_\theta(v_t)\} = 0$ and $E\{\psi_\theta(V_{jt})\} = 0$ implied by Assumption 2(vi) are first imposed for a starting value of θ, they may not be satisfied for other subsequently chosen values of θ. On the other hand, the slope estimator is consistent regardless of the value of θ. Hence, in order to propose a test for any value of θ, we use the slope estimators only to construct a test statistic (denoted by KM). Specifically, let $\alpha_{0(1)}$ and $\alpha_{0(2)}$ be the intercept and slope coefficients, respectively, and also let us decompose the quantile estimators $\tilde{\alpha}$ and $\hat{\alpha}$ accordingly; that is, $\tilde{\alpha}' = (\tilde{\alpha}_{(1)}, \tilde{\alpha}_{(2)}')$ and $\tilde{\alpha}' = (\tilde{\alpha}_{(1)}, \tilde{\alpha}_{(2)}')$. Let $R_{(2)}$ be the matrix composed of the last $(K_1 + G - 1)$ rows in R. Then we have the following Theorem.

Theorem 1. Suppose that kernel-type density estimators $\hat{h}(0)$, $\hat{f}(0)$, and $\hat{g}_i(0)$ are respectively consistent for $h(0)$, $f(0)$, and $g_i(0)$, respectively. Then, under Assumptions 1 and 2, we have

$$KM = T(\tilde{\alpha}_{(2)} - \hat{\alpha}_{(2)})[R_{(2)}\hat{C}^{-1}R_{(2)}']^{-1}(\tilde{\alpha}_{(2)} - \hat{\alpha}_{(2)}) \xrightarrow{d} \chi^2(K_1 + G - 1).$$

The result of Corollary 1 easily follows from Lemmas 1 and 2. In the next section, we examine the finite-sample performance of the proposed test by using Monte Carlo simulations.

4 Monte Carlo Simulations

The results obtained in the previous section hold in large samples. In this section, the finite sample properties in terms of the size and power of the proposed test are studied through Monte Carlo simulations. We use a simultaneous equation system composed of two equations. The first equation, which is the equation of interest, contains two endogenous variables at quantile θ and two exogenous variables including a constant. In total, four exogenous variables are present in the whole system. The structural simultaneous equation system can be written

$$B\begin{bmatrix} y_t \\ Y_t \end{bmatrix} + \Gamma x_t = U_t, \tag{12}$$

where $\begin{bmatrix} y_t \\ Y_t \end{bmatrix}$ is a 2×1 vector of endogenous variables, and x_t is a 4×1 vector of exogenous variables with the first element equal to one. The error term $U_t = \begin{bmatrix} u_t \\ w_t \end{bmatrix}$ is a 2×1 vector

of error terms. We specify the structural parameters as follows: $B = \begin{bmatrix} 1 & -0.3 \\ \delta & 1 \end{bmatrix}$ and $\Gamma = \begin{bmatrix} -1 & -0.2 & 0 & 0 \\ -1 & 0 & -0.4 & -0.5 \end{bmatrix}$. The system is over-identified by the zero restrictions $\Gamma_{13} = \Gamma_{14} = \Gamma_{22} = 0$.

We generate the error terms U_t using $\mathrm{N}(0_{2\times1}, I_{2\times2})$. Thus, we draw the second to fourth elements x_t from the normal distribution with mean $(0.5, 1, -0.1)'$, variances equal to 1 for normalization, $cov(x_{2t}, x_{3t}) = 0.3, cov(x_{2t}, x_{4t}) = 0.1$ and $cov(x_{3t}, x_{4t}) = 0.2$, where x_{2t}, x_{3t} and x_{4t} are the non-constant elements of x_t. Once x_t and U_t are generated, the endogenous variables y_t and Y_t are generated through (12). The first structural equation is

$$y_t = 0.3\ Y_t + 1 + 0.2\ x_{2t} + u_t, \tag{13}$$

where the presence of endogeneity will depend on the δ parameter in the second equation.

Note that if $\delta = 0$, there is no endogeneity at θ in (13). On the other hand, endogeneity at θ occurs if $\delta \neq 0$. Because the magnitude of δ determines the strength of endogeneity, we can use it to analyse the empirical power of the proposed test**.** We select three values for δ: 0.00, 0.60, and 1.20. For each of the values, we compute the rejection probabilities by the proposed KM test for the null hypothesis of no endogeneity at the 5 % significance level based on 1,000 replications.

The results are displayed in Table 1 for $T = 100$ and for some selected values of θ. First, it can be seen that when $\delta = 0$, the rejection probability in each case is close to the nominal 5 % level. As the value of δ increases, the rejection probability increases from 5 % for all values of θ considered. The level of our test is therefore satisfactory with normal distribution and the chosen sample size.

We now turn to power. The highest empirical power is around 20-26 % when δ is around 1.2. As we increase the sample size from 100 to 200, as shown in Table 2, the power of the test is further increased so that the rejection probability is in the range of 40-50 % when δ is around 1.2.

5 Computation Costs

In this section, we show that the Chernozhukov and Hansen test becomes infeasible as the number of endogenous variables increase, due to its computational burden. Suppose that we wish to estimate the following quantile model:

$$y_t = \beta_0(\theta) + \gamma_1(\theta)Y_{1t} + \gamma_2(\theta)Y_{2t} + ... + \gamma_G(\theta)Y_{Gt} + u_t,\ t = 1, ..., T,$$

where the $Y_{it}, i = 1, .., G$, are endogenous so that we have G endogenous variables. We assume that enough instruments for the Y_{it} are available and are collected in a vector $w_t = (w_{1t}, ..., w_{Kt})$, with $K > G$.

The first step to implement the IVQR method is to choose some prior values on the parameters corresponding to the endogenous variables $(\gamma_1(\theta), \gamma_2(\theta), ..., \gamma_G(\theta)$, in our case). Based on these prior values, a G-dimensional polyhedron (denoted by $\Gamma = \prod_{i=1}^{G} \Gamma_i$, where Γ_i is the parameter interval sufficiently large enough to include $\gamma_i(\theta), i = 1, ..., G$) is selected as a basis for a grid search. For a grid point of initial values $(\gamma_1, \gamma_2, ..., \gamma_G) \in \Gamma$, one can run a quantile regression of $y_t - \gamma_1 Y_{1t} - ... - \gamma_G Y_{Gt}$ on a constant and w_t, which corresponds to solving:

$$\min_{\beta_{0\theta}, \delta_\theta} \sum_{t=1}^{T} \rho_\theta \left(y_t - \gamma_1 Y_{1t} - ... - \gamma_G Y_{Gt} - \beta_{0\theta} - \delta_\theta w_t \right).$$

The resulting estimators for $\beta_{0\theta}$ and δ_θ are denoted as $\hat{\beta}_{0\theta}(\gamma_1, \gamma_2, ..., \gamma_G)$ and $\hat{\delta}_\theta(\gamma_1, \gamma_2, ..., \gamma_G)$. The coefficient δ_θ should be zero if w_t is a valid instrument because it must be uncorrelated at quantile θ with the error term. Therefore, one can consider estimating $(\gamma_1(\theta), \gamma_2(\theta), ..., \gamma_G)$ by minimizing a norm of as follows:

$$\min_{\gamma_1,\gamma_2,...,\gamma_G} \left\| \hat{\delta}_\theta(\gamma_1, \gamma_2, ..., \gamma_G) \right\|.$$

A simple case, used by Chernozhukov and Hansen, is just to minimize the Euclidean length of the estimator. The IVQR estimator is obtained by running the above initial quantile regression $\prod_{i=1}^{G} \kappa(\Gamma_i)$ times, where $\kappa(\Gamma_i)$ is the number of grid points in the parameter interval Γ_i. The computational time for IVQR will depend on how large the parameter polyhedron $\Gamma = \prod_{i=1}^{G} \Gamma_i$ is as well as how many grid points are used in each of the parameter intervals Γ_i. In contrast, our procedure requires running quantile regression only twice; that is: in the first-step and the second-step quantile regressions.

In order to show the difficulty of choosing the interval and the number of grid points for simulations, let us consider a typical application of quantile regression under endogeneity, with alternative estimators. Chevapatrakul et al. (2009) estimate the Taylor rule using our DSQR estimator, while Wolters (2012) uses instead the IVQR estimator for the same purpose. The prior value for the inflation responsiveness, which is here the parameter for the endogenous variable of interest, has a mean of 1.5, while it turns out to have a quite substantial variation across quantiles; 1.28 to 3.35 in Chevapatrakul et al. (2009) and 1.4 to 3.1 in Wolters (2012). Hence, Γ_i should be large enough to include these values, and have a grid precise enough to approximate them.

In our simulations for a fictitious case, we normalize Γ_i to be the unit interval [0,1] and we chose 50 for the number of grid points. Table 3 first shows the computation time for the KM test for a number of observations $T = 100$, 300 and 500, and a number of endogenous variables $G = 1$, 2, 3, 4 and 5. The used computer is a 2009 Pentium PC and the software used for the code is Matlab. Table 3 also reports the corresponding computation times for the Chernozhukov and Hansen test with a small number of grid points equal to 50 to favour the later test. The results show that the Chernozhukov and Hansen test is practically infeasible with more than 3 or 4 endogenous variables in that example. In such cases, our proposed method can still be applied with almost immediate results.

6 An Application to Food Engel Curve Estimation

In this section, we apply our endogeneity test to a model of Engel curves in the UK. The dataset is drawn from the UK Family Expenditure Survey conducted in 1995, which has been used in previous studies such as Blundell, Chen, and Kristensen (2007) and Chen and Pouzo (2009).[4] Linearizing these authors' specifications, we consider the following quantile Engel curve equation:

$$y_i = \beta_{0\theta} + \beta_{1\theta} x_{1i} + \gamma_\theta Y_i + u_i, \tag{14}$$

where y_i is food budget share of household i in 1995, x_{1i} is a dummy variable for children (i.e., $x_{1i} = 0$ if household i has no children and 1 if household i has at least one child), and Y_i is the log of total expenditure on both nondurable goods and services of household i in 1995. Variable u_i

[4] We are grateful to Xiaohong Chen for kindly providing us with the dataset for this research.

is an error term that we assume to be subject to a conditional quantile restriction. As mentioned in Blundell et al. (2007), Y_i may be either exogenous or endogenous, and this may be empirically tested and corrected if needed. Following these authors, the male log-earning of household i in 1995 is used as an instrument. Table 4 shows some summary statistics of the three main variables for the surveyed sample of 1665 households with two or fewer children.

In Table 5, we first present the conditional mean model estimates, obtained by using OLS and 2SLS, thus allowing Y_i to be either exogenous or endogenous across all quantiles. The estimates for the coefficient of total expenditure are negative, which indicates that the food share decreases as the total expenditure increases, as expected from the Engel law. The p-value of the usual Hausman test is 0.005, supporting endogeneity in the mean.

Lastly, Table 6 presents the test results applied to the model in (14) for each quantile index ($\theta = 0.1, 0.2, ..., 0.9$). Exogeneity is not rejected at low quantiles, whereas there exists evidence for the presence of endogeneity in the middle and high quantiles at the 10 % significance level. As discussed earlier, these results are consistent with the incidence of large omission errors in non-food expenditure.

The test results over a finer grid from 0.01 to 0.99, with increments of 0.01, are graphically displayed in Figure 1, in which the specific quantile area over which the conditional distribution is affected by endogeneity can be examined. It is obvious that the usual exogeneity tests based on means would fail to capture such complex endogeneity features.

7 Conclusion

In this paper, we have proposed a test of endogeneity in conditional quantile models. The test is based on the distance between two estimators, of which one is consistent only under no endogeneity at a given conditional quantile, while the other is consistent regardless of the presence of endogeneity. The derived asymptotic null distribution of the test statistic is the usual Chi-square distribution. Monte Carlo simulations indicate that the test has good size and power properties, even in finite samples. Moreover, our test can easily be used with more than a few endogenous regressors without involving infeasible computation burden as for the Chernozhukov and Hansen test. By applying the proposed test to food Engel curves estimated from UK consumption expenditure data, it is revealed that only some parts (including the median) of the conditional distribution of the food share are affected by endogeneity.

References

[1] Abadie, A., J. Angrist and G. Imbens (2002), "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91-117.

[2] Amemiya, T. (1982), "Two stage least absolute deviations estimators." *Econometrica*, 50, 689-711.

[3] Arias, O., K.F. Hallock and W. Sosa-Escudero (2001), "Individual heterogeneity in the returns to schooling: Instrumental variables quantile regression using twins data." *Empirical Economics*, 26, 7-40.

[4] Blundell, R., X. Chen and D. Kristensen (2007), "Semi-nonparametric IV estimation of shape invariant Engel curves," *Econometrica*, 75, 1613-1669.

[5] Butler, J.S. (2000), "Efficiency results of MLE and GMM estimation with sampling weights," *Journal of Econometrics*, 96, 25-37.

[6] Chen, L.-A. and S. Portnoy (1996), "Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models." *Commun. Statist.-Theory Meth.* 25, 1005-32.

[7] Chen, X., O. Linton and I. Van Keilegem (2003), "Estimation of semi-parametric models when the criterion function is not smooth," *Econometrica*, 71, 1591-1608.

[8] Chen, X. and D. Pouzo (2009), "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152, 46-60.

[9] Chernozhukov, V. and C. Hansen (2005), "An IV model of quantile treatment effects," *Econometrica*, 73, 245-261.

[10] Chernozhukov, V. and C. Hansen (2006), "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491-525.

[11] Chernozhukov, V. and C. Hansen (2008), "Instrumental variable quantile regression: a robust inference approach," *Journal of Econometrics*, 142, 379-398.

[12] Chesher, A. (2003), "Identification in nonseparable models," *Econometrica*, 71, 1405-1441.

[13] Chevapatrakul, T., T.-H. Kim and P. Mizen (2009), "The Taylor principle and monetary policy approaching a zero bound on nominal rates: quantile regression results for the United States and Japan," *Journal of Money, Credit and Banking*, 41, 1705-1723.

[14] Chmelarova, V. and R.C. Hill (2010), "The Hausman pretest estimator," *Economics Letters*, 108, 96-99.

[15] Garcia, J., P.J. Hernandez and A. Lopez-Nicolàs (2001), "How wide is the gap? An investigation of gender wage differences using quantile regression." *Empirical Economics*, 26, 149-67.

[16] Hahn, J. and J. Hausman (2002), "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, Vol. 70, N0. 1, 163-189, January.

[17] Hausman, J.A. (1978), "Specification tests in econometrics," *Econometrica*, 1251-1271.

[18] Horowitz, J.L. and S. Lee (2007), "Nonparametric instrumental variable estimation of a quantile regression model," *Econometrica*, Vol. 75, No. 4, 1191-1208.

[19] Hong, H. and E. Tamer (2003), "Inference in censored models with endogenous regressors," *Econometrica*, 71, 905-932.

[20] S.J. Jun (2008), "Weak identification robust tests in an instrumental quantile model," *Journal of Econometrics*, 144, 118-138.

[21] Kemp, G. (1999), "Least absolute error difference estimation of a single equation from a simultaneous equations system." Mimeo University of Essex.

[22] Kim, T-H. and White, H. (2003), "Estimation, inference, and specification testing for possibly misspecified quantile regression, *Advances in Econometrics*, 17, 107-132.

[23] Kim, T-H. and C. Muller (2004), "Two-stage quantile regressions when the first stage is based on quantile regressions," *The Econometrics Journal*, 7, 218-231.

[24] Kim, T-H. and C. Muller (2012), "Bias Transmission and Variance Reduction in Two-Stage Quantile Regressions," mimeo University of Aix-Marseille.

[25] Lee, S. (2007), "Endogeneity in quantile regression models: a control function approach,"*Journal of Econometrics*, 141, 1131-1158.

[26] Lee, Y. and R. Okui (2012), "Hahn-Hausman test as a specification test," *Journal of Econometrics*, 167, 133-9.

[27] Ma, L. and R. Koenker (2006), "Quantile regression methods for recursive structural equation models," *Journal of Econometrics,* 134, 471-506.

[28] Powell, J. (1983), "The asymptotic normality of two-stage least absolute deviations estimators." *Econometrica*, 51, 1569-75.

[29] Sakata, S. (2007), "Instrumental Variable Estimation Based on Conditional Median Restriction," *Journal of Econometrics*, 141, 350-382.

[30] Wolters, M.H. (2012), "Estimating monetary policy reaction functions using quantile regressions," *Journal of Macroeconomics*, 34, 342-361.

Appendix

Proof of Lemma 1: Let $\hat{\delta} = (\tilde{\alpha}', \hat{\alpha}')'$ and $\delta_0 = (\alpha_0', \alpha_0')'$. Using (9) and (10), we have

$$\begin{aligned} T^{1/2}(\hat{\delta} - \delta_0) &= \begin{bmatrix} Q_z^{-1} T^{-1/2} \sum_{t=1}^{T} Z_t \epsilon_{1t} + o_p(1) \\ Q_{zz}^{-1} H(\Pi_0)' T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_{2t} + o_p(1) \end{bmatrix} \\ &= DT^{-1/2} \sum_{t=1}^{T} S_t + o_p(1) \end{aligned} \tag{15}$$

where

$$D = \begin{bmatrix} Q_z^{-1} & 0 \\ 0 & Q_{zz}^{-1} H(\Pi_0)' \end{bmatrix} \text{ and } S_t = \begin{bmatrix} Z_t \epsilon_{1t} \\ x_t \epsilon_{2t} \end{bmatrix}.$$

Let us now consider (15). S_t is iid by Assumption 1, and $E(S_t) = 0$ under the null hypothesis of no endogeneity at θ and Assumption 2(vi). Hence, in order to apply the Lindeberg-Levy CLT to $T^{-1/2} \sum_{t=1}^{T} S_t$, it is sufficient to show that $var(S_t)$ is bounded. The moment conditions on x_t and Y_t in Assumption 2(i) are sufficient for this purpose because $\psi_\theta(\cdot)^2$ is bounded from above and all the densities evaluated at zero are bounded from below and strictly positive.

Given that

$$var(S_t) = \Omega = \begin{bmatrix} \sigma_{11} Q_z & \sigma_{12} Q_{zx} \\ \sigma_{12} Q_{zx}' & \sigma_{22} Q_x \end{bmatrix},$$

where $\sigma_{ij} = Cov(\varepsilon_{it}, \varepsilon_{jt})$, we have $T^{-1/2} \sum_{t=1}^{T} S_t \xrightarrow{d} N(0, \Omega)$, which implies that

$$T^{1/2}(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, C),$$

where $C = D\Omega D'$. Noting that $T^{1/2}(\tilde{\alpha} - \hat{\alpha}) = RT^{1/2}(\hat{\delta} - \delta_0)$, we have

$$T^{1/2}(\tilde{\alpha} - \hat{\alpha}) \xrightarrow{d} N(0, RCR'),$$

which, in turn, implies that

$$T(\tilde{\alpha} - \hat{\alpha})[RC^{-1}R']^{-1}(\tilde{\alpha} - \hat{\alpha}) \xrightarrow{d} \chi^2(K_1 + G),$$

which completes the proof. QED.

Proof of Theorem 1: Under Assumptions 1 and 2, the cross product estimators $\hat{Q}_z$, $\hat{Q}_{zx}$, and $\hat{Q}_x$ are consistent almost surely due to the Kolmogorov law of large numbers. It is also obvious to show the consistency of $\hat{Q}_{zz}$ and $H(\hat{\Pi})$ because $\hat{\Pi}$ is consistent. Hence, it remains to show the consistency of $\hat{\sigma}_{11} = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{1t} \hat{\epsilon}_{1t}, \hat{\sigma}_{22} = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{2t} \hat{\epsilon}_{2t}$ and $\hat{\sigma}_{12} = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{1t} \hat{\epsilon}_{2t}$. We provide the detailed proof only for $\hat{\sigma}_{12}$, given that the same kind of argument is used for $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$.

Recalling the definition of $\sigma_{12} = E(\epsilon_{1t} \epsilon_{2t})$, the following comes directly from the Kolmogorov law of large numbers:

$$\tilde{\sigma}_{12} - \sigma_{12} \xrightarrow{p} 0,$$

where $\tilde{\sigma}_{12} = T^{-1} \sum_{t=1}^{T} \epsilon_{1t} \epsilon_{2t}$ because (i) $\epsilon_{1t} \epsilon_{2t}$ is iid due to Assumption 1 and (ii) $E(|\epsilon_{1t} \epsilon_{2t}|) < \infty$ because of the boundedness of the $\psi_\theta(\cdot)-$function. Hence, it will be sufficient to show $\hat{\sigma}_{12} - \tilde{\sigma}_{12} \xrightarrow{p} 0$ to prove that $\hat{\sigma}_{12} - \sigma_{12} \xrightarrow{p} 0$.

Note that

$$|\hat{\sigma}_{12} - \tilde{\sigma}_{12}| \ \leq a_T + b_T + c_T,$$

where $a_T = \theta T^{-1} \sum_{t=1}^{T} |1_{[\hat{v}_t \leq 0]} - 1_{[v_t \leq 0]}|$, $b_T = \theta T^{-1} \sum_{t=1}^{T} |1_{[\hat{u}_t \leq 0]} - 1_{[u_t \leq 0]}|$

and $c_T = T^{-1} \sum_{t=1}^{T} |1_{[\hat{u}_t \leq 0]} 1_{[\hat{v}_t \leq 0]} - 1_{[u_t \leq 0]} 1_{[v_t \leq 0]}|$. The proof for both $a_T \xrightarrow{p} 0$ and $b_T \xrightarrow{p} 0$ can be found in the proof of Proposition 3 in Kim and Muller (2004). Hence we only need to show that $c_T \xrightarrow{p} 0$. We note that

$$\begin{aligned}
c_T &\leq T^{-1} \sum_{t=1}^{T} |1_{[\hat{u}_t \leq 0]}| \times |1_{[\hat{v}_t \leq 0]} - 1_{[v_t \leq 0]}| + T^{-1} \sum_{t=1}^{T} |1_{[v_t \leq 0]}| \times |1_{[\hat{u}_t \leq 0]} - 1_{[u_t \leq 0]}| \\
&\leq T^{-1} \sum_{t=1}^{T} |1_{[\hat{v}_t \leq 0]} - 1_{[v_t \leq 0]}| + T^{-1} \sum_{t=1}^{T} |1_{[\hat{u}_t \leq 0]} - 1_{[u_t \leq 0]}| \\
&\leq \theta^{-1}(a_T + b_T) \xrightarrow{p} 0.
\end{aligned}$$

Hence, the proof is completed. QED.

Table 1. Rejection probabilities by the KM test for the null hypothesis of no endogeneity at θ^{th} quantile with T = 100

	δ	θ		
		0.25	0.50	0.75
Size	0.00	0.05	0.04	0.06
	0.60	0.11	0.09	0.10
Power	1.20	0.26	0.20	0.23

Table 2. Rejection probabilities by the KM test for the null hypothesis of no endogeneity at θ^{th} quantile with T = 200

	δ	θ		
		0.25	0.50	0.75
Size	0.00	0.06	0.05	0.06
	0.60	0.31	0.29	0.31
Power	1.20	0.53	0.57	0.52

Table 3. Computational times when the model has G endogenous variables and T observations

			Number of endogenous variables (G)				
			1	2	3	4	5
KM Test (Time *in Seconds*)	Number of observations (T)	100	0.1	0.06	0.06	0.1	0.1
		300	0.16	0.18	0.22	0.32	0.46
		500	0.18	0.38	0.54	0.82	0.84
C&H test (Time *in Hours*)	Number of observations (T)	100	0.0007	0.0215	1	82	4,080
		300	0.0011	0.0646	4	271	20,313
		500	0.0013	0.1299	9	703	36,545

Table 4. Summary statistics

	mean	Std.
Food share	0.2074	0.0971
Log expenditure	5.4215	0.4494
Log earnings	5.8581	0.5381
Sample size	1665	

Table 5. Conditional mean regression

	Estimates (Standard Errors)		
	Intercept	Dummy for Kids	Expenditure
OLS	0.761 (0.024)	0.056 (0.004)	-0.109 (0.004)
2SLS	0.614 (0.047)	0.054 (0.004)	-0.081 (0.009)
Hausman Test	Statistic = 13.02 P-value = 0.005.		

Table 6. Test for endogeneity at various quantiles

Quantile (θ)	Test Statistic (KM)	P-value
0.1	4.028	0 .133
0.2	2.902	0.234
0.3	2.582	0.275
0.4	11.829	0.003
0.5	4.742	0.093
0.6	4.952	0.084
0.7	9.362	0.009
0.8	7.587	0.023
0.9	7.898	0.019

Figure 1. P-values of the KM test over a quantile grid of [0.01, 0.99]